Using Data-driven and Phonetic Units for Speaker Verification

Asmaa El Hannani[1,2]*, Doroteo T. Toledano*[3]*, Dijana Petrovska-Delacrétaz*[2]*,*
Alberto Montero-Asenjo[3] *and Jean Hennebert*[1]

[1]DIVA Group, Informatics Dept., University of Fribourg, Switzerland
{*asmaa.elhannani, jean.hennebert*}*@unifr.ch*
[2]EPH Dept., Institut National des Télécommunication, Evry, France
dijana.petrovska@int-evry.fr
[3]ATVS, Universidad Autonoma de Madrid, Spain
{*doroteo.torre, alberto.montero*}*@uam.es*

Abstract

Recognition of speaker identity based on modeling the streams produced by phonetic decoders (phonetic speaker recognition) has gained popularity during the past few years. Two of the major problems that arise when phone based systems are being developed are the possible mismatches between the development and evaluation data and the lack of transcribed databases. Data-driven segmentation techniques provide a potential solution to these problems because they do not use transcribed data and can easily be applied on development data minimizing the mismatches. In this paper we compare speaker recognition results using phonetic and data-driven decoders. To this end, we have compared the results obtained with a speaker recognition system based on data-driven acoustic units and phonetic speaker recognition systems trained on Spanish and English data. Results obtained on the NIST 2005 Speaker Recognition Evaluation data show that the data-driven approach outperforms the phonetic one and that further improvements can be achieved by combining both approaches.

1. Introduction

In recent years, research on text-independent speaker verification has expanded from using only the acoustic content of speech to trying to utilise high-level information, such as linguistic content, pronunciation and idiolectal word usage. Works examining the exploitation of high-level information sources have provided strong evidence that gains in speaker recognition accuracy are possible [1]. In [2], the possibility of using word n-gram statistics for speaker verification is explored. This technique exploits the idiolectal information in a straightforward way and gave encouraging results. Motivated by the work of [2], similar techniques have been applied to phone n-gram statistics [3]. This last approach gave good results and was found to provide features complementary to short-term acoustic features. These promising techniques are however built using manually transcribed databases that are error-prone and expensive to create. These databases need also to be updated with new data sets in order to match with potentially new specifications (channel, microphones, context of use, ...) of the verification data.

In [4] we presented a similar system to [3] but we used the automatic segmentation based on Automatic Language Independent Speech Processing (ALISP) tools [5] instead of the phonetic one. In this system, speaker specific information is captured only by analyzing sequences of ALISP units. The main advantage of such systems is that the ALISP-sequences are automatically acquired from the raw speech data with no need of manually transcribed databases. As a result, the ALISP recognizer can be easily trained on speech data matching closely the acoustic conditions for any target application.

The work described in this paper has concentrated on comparing two speaker verification systems, the first one based on data-driven ALISP units and the second one on phonetic units. Through this work, our objective is to determine if data-driven ALISP units can be used to substitute or complement phonetic units and under which conditions. Another important aspect that we attempted to analyze is the correlation between the automatically aligned phonemes and ALISP units. Somehow linked to this correlation aspect is the question wether the ALISP system brings different speaker-specific information than the phonetic system and can therefore provide improvements when fused with it.

The outline of this paper is the following: Section 2 describes the phonetic and the ALISP data-driven systems. In section 3 the evaluation results are reported. The conclusions are given in section 4.

2. Description of the Experiments

We present in this section the experimental setup of the phonetic and ALISP data-driven systems. Both systems include two parts. The first one consists in building the decoder that outputs labeling of the speech data. The second part consists in performing speaker verification, based on these labelings. The phonetic and ALISP decoders (sections 2.1 and 2.2) are built independently at different sites, resulting in the use of different front-end processing. Whereas the speaker recognition task (sections 2.3 and 2.4) is carried out on a common protocol using the same modeling methods and the same databases.

2.1. Training of the Phonetic HMMs

The acoustic processing for the phonetic system is based on the Advanced Distributed Speech Recognition Front-End [6] of the European Telecommunications Standards Institute (ETSI). This standard defines a distributed speech recognition front-end intended to reside in mobile terminals so that the speech recognition features (instead of the coded speech) are transmitted over the network to a server where the speech recognizers run. The

front-end is based on the standard Mel Frequency Cepstral Coefficients (MFCCs), including 13 static features (C0 to C12), deltas and double deltas for a total of 39 components. Additionally, it includes mechanisms for robustness against channel distortion (blind equalization) and additive noise (double Wiener filter).

The phonetic decoders are based on Hidden Markov Models (HMMs) and implemented using HTK [7]. The phonetic HMMs are three-state left-to-right models with no skips, and the output pdf of each state modeled as a weighted mixture of Gaussians. Phonetic HMMs are trained for Castilian Spanish and American English using the Albayzin [8] and TIMIT [9] corpora. Since these corpora are microphone corpora sampled at 16 kHz, we filtered them to simulate a telephone channel and then downsampled them to 8 kHz before training and testing the models. Context-independent phonetic HMMs are used. 23 phones are considered for Castilian Spanish and 39 for American English (the phoneme set of the the CMU pronouncing dictionary [10]). We trained phonetic HMMs with 1 to 80 mixtures per state for each of the two languages and selected the number of mixtures per state that performed best in a speaker verification test using a subset of NIST SRE 2004 data. The best performance was achieved with just 3 mixtures per state in English and 15 in Spanish [11]. All phonetic results presented in this paper were obtained with those models.

2.2. Training of the ALISP Data-driven HMMs

The speech parameterization is done with Mel Frequency Cepstral Coefficients (MFCCs), using HTK [7]. Mel frequency bands are computed in the 300-3400 Hz range. Cepstral mean substraction is applied to the 15 static coefficients, estimating the mean on the speech-detected parts of the signal. The energy and delta components are appended, leading to 32 coefficients in each feature vector.

The data-driven speech units, denoted here as ALISP units, are automatically determined from the training corpus, with no need of phonetic transcription of the corpus. The steps needed to acquire and model the ALISP units are the following. After the pre-processing step, temporal decomposition [12] is used to obtain an initial segmentation of the speech data into quasi-stationary segments. The speech segments correspond actually to spectrally stable portions of the signal. We then compute the gravity center for each segment and train a gender dependent vector quantizer to cluster these centers of gravity. The codebook size (65 in our case) defines the number of ALISP symbols. The initial labeling of the entire speech segments is achieved using minimization of the cumulated distances of all the vectors from the speech segment to the nearest centroid of the codebook. The result of this step is an initial segmentation and labeling of the training corpus. Using HTK, Hidden Markov Models (HMMs) are then initialized from this labeled segments to build a set of 65 ALISP units. The HMM units are then re-trained on the data set using an ergodic topology and applying a Baum-Welch re-estimation. Each ALISP unit is modeled by a left-to-right HMM having three emitting states and containing up to 8 Gaussians each. The number of Gaussians is determined through a dynamic splitting procedure. The number of ALISP classes, 65, was chosen in order to have a comparable segmentation with any phonetic transcription. More experiments will be done by varying this number. The gender dependent ALISP HMMs are trained on data from (1999, 2001 and 2003) NIST SRE data sets.

2.3. Development and Evaluation Databases

In order to achieve comparable results an experimental setup was agreed between the sites developing this work. The experimental setup defined a similar technique of using the phonetic and pseudo-phonetic labelings obtained by the phonetic and ALISP decoders, as well as the development and evaluation databases that should be used to train and evaluate the resulting speaker recognition systems. In particular, we decided to use the complete NIST 2004 SRE corpora to develop the speaker recognition systems based on the phonetic and ALISP decoders and the 8conv4w-1conv4w evaluation condition of the NIST 2005 SRE corpora to evaluate and compare results. Next sections describe how these corpora were used and the techniques used to model phonetic and ALISP sequences and perform speaker recognition based on these models.

2.4. Speaker Recognition Based on Phonetic and Pseudo-phonetic Labeling

The phonetic and ALISP HMMs described in previous sections are used to produce phonetic and pseudo-phonetic labelings of the development and evaluation databases.

2.4.1. Training of the Language Models

The label sequences produced by the phonetic and data-driven recognizer are used to train phone and ALISP trigrams using the HTK LM tools. The trigrams construction is a tow stage process. Firstly, the training text is scanned and the trigrams are counted and stored in a database of gram files. Secondly the resulting gram files are used to compute trigram probabilities which are stored in the language models file. The trigram language models is used to predict each symbol in the sequence given its tow predecessors. During the testing phase, the label sequences of a previously unseen test text is scored against the language model (see 14^{th} chapter of the HTK book [7] for more details). The speaker specific language models are adapted from a Universal Background Models (UBMs). The UBMs are trained on all NIST 2004 SRE training data. Then, speaker models SM_i are adapted from the UBMs using the 8-conversations available for training in the NIST 2005 SRE corpus (8conv4w-1conv4w condition in the evaluation).

It is important to emphasize that the methods used to train the language models are basically the same for phonetic and ALISP units with only a couple of differences: first the adaptation factor used, which is different for each decoder (in part due to the different number of phonetic or pseudo-phonetic units used) and second the use of a gender-dependent UBM in the case of ALISP units.

2.4.2. Scoring

Given a test utterance, we first produce its labeling $X = x_1, ..., x_{N_p}$ using the phonetic and ALISP decoders in the same way as when training the models. The sequence of labels x_i are then used to compute the likelihoods with the statistical models SM_i and UBM:

$$\begin{aligned} L_{Si} &= P(X|SM_i) \\ L_U &= P(X|UBM) \end{aligned} \tag{1}$$

Our recognition score is a log-likelihood ratio computed with:

$$Score_i = \frac{1}{N_p}\, log\left[\frac{L_{Si}}{L_U}\right] \tag{2}$$

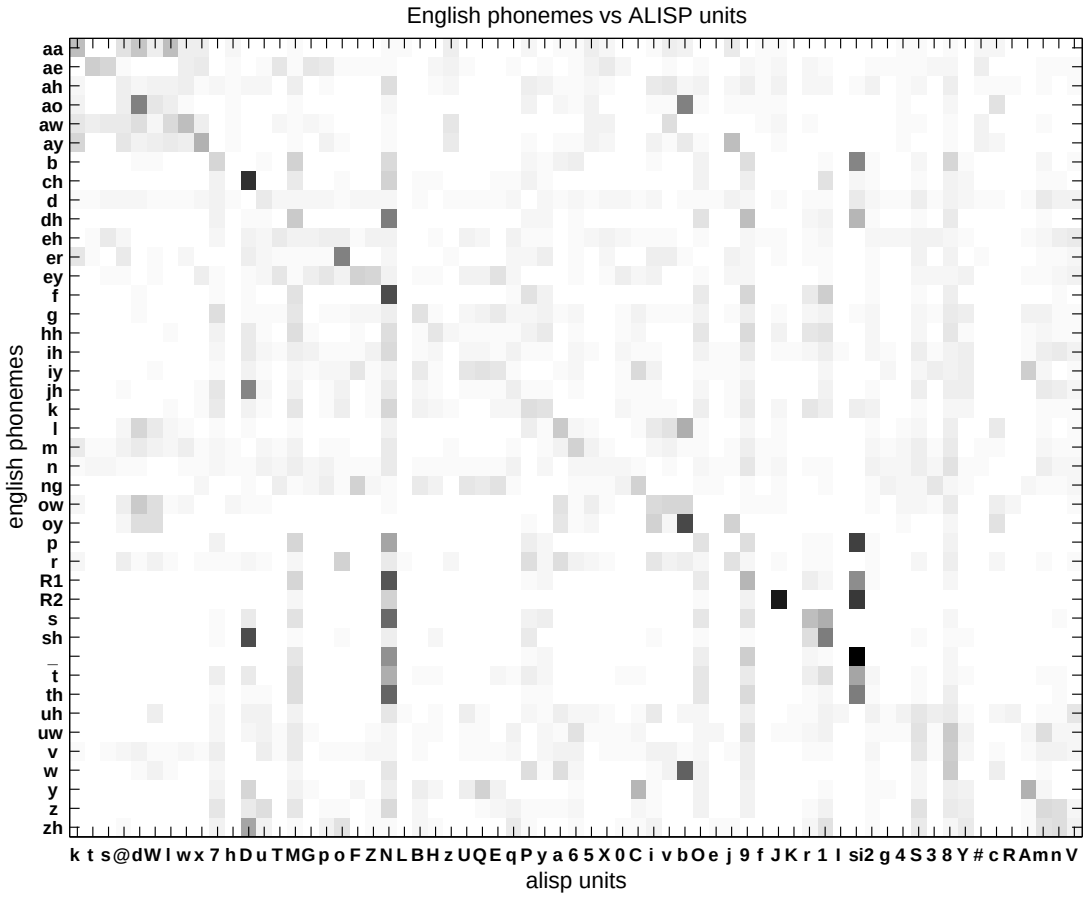

Figure 1: Comparison of the correspondence between the ALISP units and the English phonemes on 500 test files. The black color corresponds to the maximum correlation and the white color to a null correlation. Gray levels are applied linearly with the value of alignment correlation. Symbols used for ALISP units are just an enumeration from 1 to 9, a to z and A to Z.

where N_p is the number of phonetic or ALISP symbols in the test utterance. Again, it must be noted that the method used for scoring with phonetic and ALISP units was exactly the same.

2.5. Systems Fusion

In this work, a Multi-Layer Perceptrons (MLP) [13] is used for the systems fusion. This perceptron has a layer consisting of inputs for each fused system, a hidden layer with 5 neurons, and an output layer using sigmoid as activation function. The MLP is gender independent and is trained on the developement data set.

3. Results and Discussion

3.1. Phonemes vs ALISP Units

We first report some characteristics about the phonetic and the ALISP labelings. Table 1 shows the results of statistics computed on 500 test files. The mean length of the ALISP units is smaller than the mean length of the English and Spanish phonemes. This is due to the fact that we have more ALISP classes than phonemes. We have also compared the correspondence between the data-driven and phonetic labelings. To quantify this correspondence we measured the overlapping between the ALISP units and the phonetic units:

$$o_{i,j} = \frac{\sum_{k=1}^{K_i} O(p_{ik}, a_j)}{\sum_{k=1}^{K_i} L(p_{ik})}$$

where $O(p_{ik}, a_j)$ represents the absolute overlapping between the k^{th} occurrence of the phoneme p_i and the ALISP unit a_j. $L(p_{ik})$ is the length of k^{th} occurrence of the phoneme p_i. The sums are over all of the K_i occurrences of p_i in the data. The value $o_{i,j}$ measures then the correspondence between the i^{th} phoneme and the j^{th} ALISP unit.

Figure 1 shows the confusion matrix built using the values $o_{i,j}$. The black color represents a maximum correlation of 1. While the confusion matrix is difficult to interpret because ALISP units outnumber English phonemes, we can conclude that there is a limited correlation between alignments of ALISP units and phonemes. On Figure 2, the confusion matrix between the Spanish and English phonemes shows also a limited correlation. Here, the reason is probably linked to the fact that English and Spanish phonemes are by nature less correlated than, for example, French and Spanish.

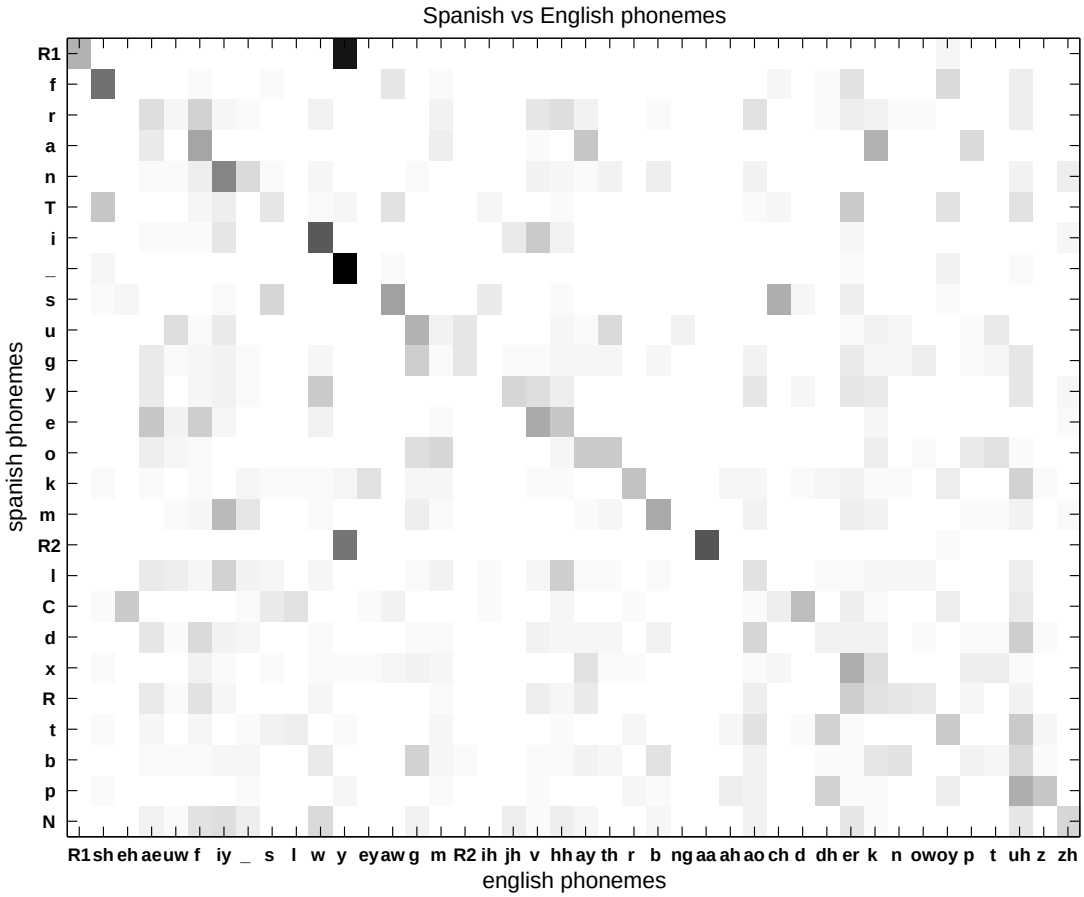

Figure 2: Comparison of the correspondence between the Spanish and the English phonemes on 500 test files. The black color corresponds to the maximum correlation and the white color to a null correlation.

Decoder	# labels classes	# labels	Average lenght (sec)
Spanish	26	966,739	0.16
English	42	1,082,717	0.14
ALISP	65	1,220,886	0.12

Table 1: Comparison of the Spanish, English and ALISP labelings.

3.2. Speaker Recognition

The second set of experiments concerns the comparison of the individual performances of the phonetic and ALISP data-driven systems. They are evaluated on the 8conv-1conv task data of the NIST2005 Speaker Recognition Evaluation campaign (with roughly 40min of speech available to train the client models). Performance is reported in term of the Detection Error Tradeoff (DET) curve [14]. Results are compared via Equal Error Rates (EER): the error at the threshold which gives equal miss and false alarm probabilities. Figure 3 shows the phonetic speaker recognition performance using the English and the Spanish decoders. The Spanish decoder is performing slightly better than the English system. Figure 3 shows also a MLP fusion of scores from the two languages that provides better results than the individual systems on their own. This result indicates that the two phonetic systems do contain complementary information which confirm the results obtained in [15] and is in accordance to the confusion matrices presented in section 3.1, which showed only limited correlation between the labelings obtained with the different phonetic decoders.

The development of the phonetic decoders used to obtain the results shown in figure 3 required the use of two phonetically transcribed corpora, one in Castilian Spanish and another in English. The availability of phonetically transcribed corpora could be an issue, specially if this approach is to be extended to a number of different languages. Moreover, the characteristics of the corpora used to train the phonetic models do not match the characteristics of the working conditions of the speaker recognition systems. In our case, the corpora used to train the phonetic decoder included 16 kHz microphone read speech, while the evaluation conditions included 8 kHz telephone spontaneous speech. To obtain the results attained in this work the conditions of the training corpora had to be adapted to be more close to the evaluation conditions (at least matching speech bandwidth).

An alternative approach that solves these two problems is using data-driven phone-like units derived directly from untranscribed speech. This way the availability of corpora is much less an issue and the training corpus can be chosen to match the working conditions as much as possible. In this work we com-

pare this second alternative (exemplified by ALISP data-driven units) to the more traditional alternative of using language-dependent phonetic units.

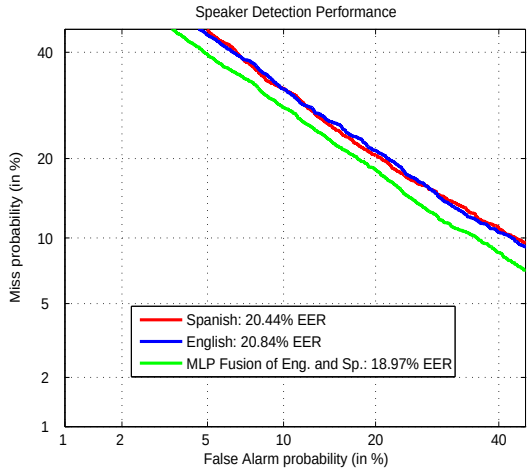

Figure 3: DET plot showing the performance of the English and Spanish phonetic speaker recognition systems and of their fusion.

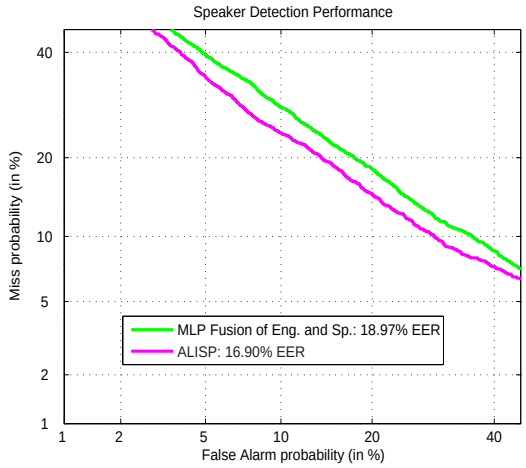

Figure 4: DET plot showing the performance of the phonetic and data-driven based speaker recognition systems.

Figue 4 shows the performance of the ALISP data-driven units and compares it to the performance of the phone based system. The EER of the ALISP data-driven system is 16.90% compared to 18.97%, the best performance of the phonetic system (MLP fusion of the English and Spanish systems). The data-driven based approach outperforms the phonetic one. These results suggest that the data-driven system may yield more robust estimation of the speaker characterization than the phonetic decoding. One possible explanation to this result is the mismatch between the training data of the phonetic decoders and the evaluation data. Note that to train the ALISP decoder we used data from (1999, 2001 and 2003) NIST SRE data sets, which is closer to NIST 2005 SRE data. Hence we can conclude that better speaker recognition results can be achieved using data-driven units than phonemes, at least if there is little or no transcribed data available recorded in similar conditions as the evaluation data.

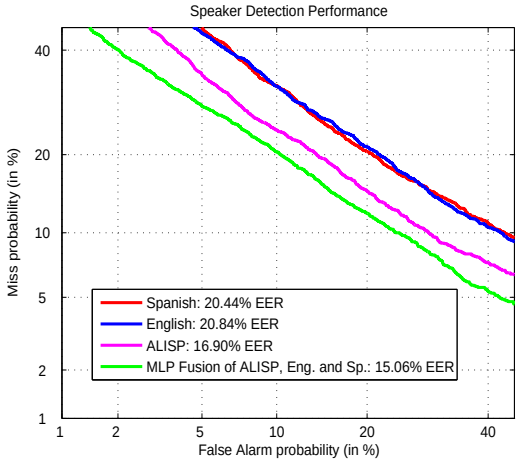

Figure 5: DET plot showing the performance of the phonetic and data-driven based speaker recognition systems and of their fusion.

As a final test, we performed a MLP based fusion of the results obtained with the ALISP and the phonetic systems. Figure 5 shows that results improve even more when fusing these systems, reaching an EER of 15.06%. This, again, is consistent with the observation made in the confusion matrices (section 3.1) of a weak correlation between the data-driven ALISP labeling and the English and Spanish phonetic labelings.

4. Conclusions

This paper has provided a comparison between two text independent speaker verification systems, one using phonemes and the other data-driven units automatically acquired from the speech data. The systems were evaluated on 8-conversations training task from NIST 2005 Speaker Recognition Evaluation. The ALISP data-driven approach achieves an EER of 16.90% compared to 18.97% for the fused phonetic systems. This means that data-driven units can substitute phonetic units for speaker recognition providing better results and with the additional advantage of not requiring phonetically transcribed speech. On the other hand, fusing the three systems further improves speaker recognition results, reaching an EER of 15.06%. This shows that data-driven units can not only substitute but also complement phonetic units to reach better speaker recognition performances.

The comparison of the labelings attained by the three different decoders shows that there is only limited correlation between the labelings produced. While this limited correlation might be considered a problem from the point of view of the consistency of the phonetic or pseudo-phonetic decoding, it seems to be positive for speaker recognition as it translates into notable gains when fusing the scores produced by the different systems.

5. References

[1] D. Reynolds, W. Andrews, J. Campbell, J. Navratil, B. Peskin, A. Adami, Q. Jin, D. Klusacek, J. Abramson, R. Mihaescu, J. Godfrey, J. Jones, and B. Xiang, "The supersid project: Exploiting high-level information for high-accuracy speaker recognition," *In Proc. ICASSP*, April 2003.

[2] G. Doddington, "Speaker recognition based on idiolectal differences between speakers," *Eurospeech*, vol. vol. 4, pp. 2517–2520, 2001.

[3] W. Andrews, M. Kohler, J. Campbell, and J. Godfrey, "Phonetic, idiolectal, and acoustic speaker recognition," *Speaker Odyssey Workshop*, 2001.

[4] A. E. Hannani and D. Petrovska-Delacrétaz, "Exploiting high-level information provided by alisp in speaker recognition," *Non Linear Speech Processing Workshop (NOLISP 05)*, 19-22 April 2005.

[5] G. Chollet, J. Černocký, A. Constantinescu, S. Deligne, and F. Bimbot, "Towards ALISP: a proposal for Automatic Language Independent Speech Processing," *In Keith Ponting, editor, NATO ASI: Computational models of speech pattern processing Springer Verlag*, 1999.

[6] ETSI ES 202 050 (v1.1.3): Speech processing, transmision and quality aspects (STQ); distributed speech recognition;advanced front-end features extraction algorithm;compression algorithms.

[7] S. Young. Hidden markov model toolkit (HTK). [Online]. Available: http://htk.eng.cam.ac.uk/

[8] A. Moreno, D.Poch, A.Bonafonte, E.Lleida, J.Llisterri, and C. J.Marino, "ALBAYZIN speech database: Design of the phonetic corpus," *Proc. Eurospeech'93*, vol. vol. 1, pp. 175–178, 1993.

[9] TIMIT acoustic-phonetic continuous speech corpus. [Online]. Available: National Institute of Standards and Technology Speech Disc 1-1.1,NTIS Order No.PB91-5050651996

[10] The CMU pronouncing dictionary. [Online]. Available: http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[11] D. T. Toledano, C. Fombella, J. Gonzalez-Rodriguez, and L. Hernndez, "On the relationship between phonetic modeling precision and phonetic speaker recognition accuracy," *In proc. of InterSpeech*, September 2005.

[12] B. Atal, "Efficient coding of LPC parameters by temporal decomposition," *Proc. ICASSP*, pp. 81–84, 1983.

[13] S. Haykin, *Neural Networks: A Comprehensive Foundation*. IEEE Computer society Press, 1994.

[14] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The det curve in assessment of detection task performance," *Proc. Eurospeech'97*, vol. vol. 4, pp. 1895–1898, 1997.

[15] W. M. Campbell, J. P. Campbell, D. Reynolds, D. A. Jones, and T. R. Leek, "Phonetic speaker recognition with support vector machines," *In Proc. Neural Information Processing Systems Conference in Vancouver*, pp. 361–388, 2003.